The Canadian Journal for the Scholarship of Teaching and Learning

Volume 13 | Issue 1 Article 6

Winter 02-15-2022

A Data-First Approach to Learning Real-World Statistical Modeling

Luke Bornn
Simon Fraser University, lbornn@sfu.ca
Jacob Mortensen
Simon Fraser University, jmortens@sfu.ca
Daria Ahrensmeier
Simon Fraser University, dahrensm@sfu.ca

https://doi.org/10.5206/cjsotlrcacea.2022.1.10204

Recommended Citation
Bornn, L., Mortensen, J., & Ahrensmeier, D. (2022). A data-first approach to learning real-world statistical modeling. *The Canadian Journal for the Scholarship of Teaching and Learning, 13*(1).
https://doi.org/10.5206/cjsotlrcacea.2022.1.10204

A Data-First Approach to Learning Real-World Statistical Modeling

Abstract

This paper presents a novel design for an upper-level undergraduate statistics course structured around data rather than methods. The course is designed around curated datasets to reflect real-world data science practice and engages students in experiential and peer learning using the data science competition platform Kaggle. Peer learning is further encouraged by patterning the course after a genetic algorithm: students have access to each other's solutions, allowing them to learn from what others have done and figure out how to improve upon previous work from week to week. Implementation details for the course are provided, and course efficacy is assessed using a survey of students and a focus group. Student responses suggest that the structure of the course contributed to narrowing the perceived gap between low- and high-performing students, that desired learning outcomes were successfully achieved, and that a data-first approach to learning statistics is effective for learning.

Cet article présente une conception nouvelle pour un cours de statistiques de fin de premier cycle structuré autour de données plutôt que de méthodes. Le cours est conçu autour d'ensembles de données conservées qui reflètent les pratiques de la science des données du monde réel et engage les étudiants et les étudiantes dans un apprentissage basé sur l'expérience et sur l'apprentissage par les pairs en utilisant la plate-forme de compétition de science des données Kaggle. L'apprentissage par les pairs est également encouragé grâce à la modélisation du cours sur un algorithme génétique : les étudiants et les étudiantes ont accès aux solutions des uns et des autres, ce qui leur permet d'apprendre à partir de ce que les autres ont fait et de trouver comment s'améliorer par rapport à leur travail passé d'une semaine à l'autre. Les détails de la mise en oeuvre du cours sont fournis et l'efficacité du cours est évaluée grâce à un sondage auprès des étudiants et des étudiantes et à un groupe de discussion. Les réponses des étudiants et des étudiantes suggèrent que la structure du cours a contribué à réduire l'écart perçu qui existait entre les étudiants et les étudiantes les plus forts et les plus faibles, que les résultats d'apprentissage désirés avaient été atteints et que l'approche axée sur les données pour l'apprentissage des statistiques était efficace pour l'apprentissage.

Keywords

statistics education research, data science learning, authentic assessment; recherche sur l'enseignement des statistiques, apprentissage de la science des données, évaluation authentique

This research paper/Rapport de recherche is available in The Canadian Journal for the Scholarship of Teaching and Learning: https://doi.org/10.5206/cjsotlrcacea.2022.1.10204

Applied statistics courses in post-secondary schools/undergraduate programs are frequently taught as a collection of data analysis techniques. For example, a standard linear regression course, which might appear in a statistics undergraduate curriculum following several introductory courses, often follows the approach of "here is method X, now apply method X to dataset Y," where dataset Y is cleaned and perfectly polished in preparation for method X. While this approach is sensible from the perspective of class organization, in many ways it is not reflective of the experience of a working statistician. Real-world industrial applications start with a problem and data (or a problem and a need to collect data) with no clear direction on methods. In industry, statisticians are often expected to work with a huge, unstructured collection of data representing complex phenomena, with the underlying task to select and develop methods to solve a specific problem. As an example from the first author's previous work experience, the statistician may be asked to identify patterns of purchase behavior from the unstructured transaction logs arising from Electronic Arts' FIFA video game. Essential skills in this setting include the ability to think critically about data; to understand when methods known to the data scientist are appropriate, and when they are not; and to learn about alternative analysis approaches. Courses focused on methods rather than data are poorly equipped to help students develop such skills. Additionally, students are clearly aware of the actual requirements of working as a statistician. In a recent study of post-SAT students, 49% of respondees classified statistics as the "analysis and interpretation of data" rather than alternatives such as "using a collection of techniques" (19%) (Bond et al., 2012, p. 14).

The use of meaningful, rather than simulated, data in statistics courses has been shown to have strong cognitive and motivational benefits (Fawcett, 2017; Neumann et al., 2013; Scheaffer, 2001), with many handbooks advocating a stronger use of authentic data in learning strategies (Dunn et al., 2012; Garfield & Ben-Zvi, 2009; Singer & Willett, 1990). In fact, there exists a significant literature on improving statistical education with various strategies, including context-sensitive, data-centered approaches to learning statistics (Cobb & McClain, 2004; Makar & Ben-Zvi, 2011; Scheaffer, 2001; Stander & Dalla Valle, 2017). The latter reference, for example, outlined a course at the University of Plymouth employing real-world data (including social media data) to engage students in learning data science and the resultant positive student outcomes.

With the design of the course "Learning from Big Data," we act upon these recommendations, making data the course focus, rather than a supporting element. From a constructivist perspective, this data-first strategy is intended to help students construct knowledge through their hands-on work with real data sets (Garfield & Ben-Zvi, 2009), with the instructor providing just enough guidance to assist in the students' understanding and subsequent modeling choices (Neumann et al., 2013). The primary intent, therefore, is that students' statistical understanding will grow through an interplay between their contextual knowledge of the problem at hand and their emerging knowledge of statistical methods (Dierdorp et al., 2011; Pfannkuch, 2011). This course shares some similarities with the capstone course described by Lazar et al. (2011), which also takes a data-first approach, but our course differs by including both a competition aspect and encouraging peer learning by explicitly sharing student solutions with the class.

In addition to this data-first approach, we further encourage experiential learning by borrowing ideas from the implementation of genetic algorithms (Holland, 1984). Genetic algorithms are an optimization technique in which parents (proposed values) are evaluated according to some objective function, then traits of the parents are combined and passed on to their children (new proposed values) in a crossover step, with parents who performed better according to the function passing on their traits with greater probability. This crossover step is followed by a mutation, in which proposed values are randomly altered, at which point the children are evaluated according to the objective, becoming the new parents, and the process is repeated until convergence.

This class followed a similar pattern. Each week students competed to develop methods that performed best on some metric, typically predictive accuracy. At the end of each week, all solutions were made available to the class, and because a single data set was analyzed over multiple weeks, students were able to borrow ideas from the best performing methods and "mutate" them to create new solutions. Iterating in this way allowed students not only to get experience working directly with data, but also exposed them to the methods and approaches fellow students used to tackle the same problems, and subsequently how each of those methods performed.

In this paper, we present the design of the course, its implementation at the undergraduate level, and our post-course analysis undertaken to study student perceptions of learning at Simon Fraser University. We conclude with a discussion of our findings and implications for future implementations.

Course Design

Rationale

In most courses in the areas of statistics and machine learning, focus is on teaching methods or theory, with applications introduced to give students hands-on experience after studying the methods. In contrast, data science in the real world proceeds in a different order, starting with data and a problem to solve, and the practitioner must decide which methods to apply, as well as how to apply them. This "decide which methods to apply" stage is frequently absent from statistics courses. It is this omission which the course "Learning from Big Data" seeks to eliminate, as it is data-centered with an emphasis on experiential learning. Significantly, in this course, students are not expected to draw upon lectures or course notes to accomplish their tasks, but rather to rely on previous knowledge, seek out external resources, and learn from their peers, which are all less formal avenues for learning that more accurately represent the experience of a practicing data scientist.

Targeted at advanced undergraduate students in computer science and statistics, the course's purpose is to equip them with the tools and experience to handle large-scale, real-life data. Though some of these students may pursue graduate studies after completion of their degrees, the majority immediately enter the workforce. By working directly with real-world datasets taken from open sources and industry collaborations, the students learn technical skills crucial for career success.

Learning Goals

The semester-long course is based around four modules of approximately equal length, where in each module, students are faced with a massive, real-world dataset. They compete in small groups to understand and model the data (including choices of methods) along some instructor-defined objective, such as classifying a binary outcome. Modules (and hence datasets) are focused on the areas of regression, classification, time series, and text analysis, though the exact splits could be tailored to specific audiences, for example health data for biostatistics students.

The learning outcomes for the course are:

1. Given a data set and accompanying problem, students will be able to identify a set of suitable (and rule out unsuitable) methods for the task at hand.

2. Students will be able to apply a broad suite of tools to data science problems.
3. Students will be able to collaborate with peers towards a common objective.
4. Students will be able to write documentation that allows others to reproduce and expand their work.
5. Students will be able to evaluate the strengths and limitations of new statistical methods and be comfortable experimenting with them.

From the perspective of learning theory and corresponding instructional design, the course design focuses on constructivism and authentic assessment. Constructivism in a variety of forms has become a key perspective for teaching science since the 1990s, informed by (and informing) science education research (see Duit, 1996). It has motivated changes towards, for example, active learning, or problem-based learning (see Tobin, 1995 for a collection of perspectives).

The "data first" approach of this course also builds upon the concept of authentic practice (e.g., Dierdorp et al., 2011) and adds intentionally designed authentic assessment, as defined by the critical elements outlined in the paper by Ashford-Rowe et al. (2014). Specifically, the assessment is challenging (incomplete data sets), the outcome is a product (model/prediction on withheld data), it requires transfer of knowledge (methods learned in other contexts), and it includes metacognition (reflection on performance and reasons for result). Additionally, the course simulates and measures a real-world test of ability (competition) within a real-world environment (Kaggle), and requires collaborative learning (within each group and even between groups) – thus incorporating several aspects of constructivism as applied to learning science.

Class Format and Assessment

Here we describe how the class is structured to achieve these learning goals and how it is motivated by constructivist philosophy. Rather than receive formal lecture-style instruction, students are encouraged to construct new knowledge by engaging with data and questions that expose gaps in existing knowledge and then learn how to fill them. At the outset of each module, students are randomly assigned to teams of 2 or 3 (randomly, but to prevent duplicate partners across modules). In the first week of each module, all teams are given the training subset of the data and conduct an exploratory data analysis, demonstrating that they understand the problem being posed and highlighting interesting features in the data that might potentially be used to solve that problem. In each subsequent week of the module every team submits their predictions on a withheld, testing portion of data.

Hence, while the course is focused on only a handful of datasets, the progress of learning is at the week to week level, where students work in groups to model and understand the dataset, with the aim of predicting some feature of a held out test set. As an example, the dataset might consist of 10,000 product purchases, including customer information. Given 60% of the data (the training set), groups would work to build a model of purchases as a function of customer information, then subsequently predict purchases on the remaining 40% of data, upon which students have customer information, but not purchase details.

At the end of each week, every student team enters their submission in an online tool, in our case the Kaggle data science platform, where it is graded on a quantitative metric (such as mean squared error, F1 score, or some other loss function). The teams are ranked each week based on their performance on the withheld data, resulting in a weekly leaderboard showcasing each team and their performance. Each week, all teams distribute a short description of their prediction method with accompanying code and results to all other teams through a push to a shared github repository. Accordingly, groups share code/information to learn from one another's progress, making it a true peer-learning environment.

Allowing students to learn in this way ties neatly with this description of how knowledge is constructed by Driver et al. (1994):

> From this perspective knowledge and understandings, including scientific understandings, are constructed when individuals engage socially in talk and activity about shared problems or tasks … learning is seen as the process by which individuals are introduced to a culture by more skilled members. (p. 7)

By engaging with the data first and seeking to answer an analysis question, students learn why techniques they are familiar with may or may not be suitable for this specific task. This learning takes place through a dialogue, first with their group, potentially with external sources on the Internet or in textbooks, with the broader class, and with the instructor. The effect of having teams see each others' code and performance is that students are able to learn from their peers to see what methods or techniques were successful, and why.

Driver et al. (1994) also highlight the importance of the instructor in learning, stating that "If students are to adopt scientific ways of knowing, then intervention and negotiation with an authority, usually the teacher, is essential" (p. 11). In this class, the instructor acts as a guide, discussing why certain techniques may or may not be appropriate and interrogating students' statistical model choices through conversation and discussion. Crucially, instructors never explicitly endorse one method as the "right" technique, instead accomplishing the first learning objective by allowing students to learn through experience which methods were appropriate. This leads to teams adopting ideas of the top-performing teams, then subsequently diverging into new ideas to further improve performance. In this way the course functions like a genetic algorithm, where low performing teams learn from (and adopt the methods of) better performing teams each week, and subsequently evolve from there.

In addition to the foundational elements of the course as described above, a series of micro-lectures is included as well to introduce a broad array of topics that may not typically be covered in a statistics curriculum. This includes, for example, version control, model averaging, missing data methods, and modeling techniques. While inclusion of these micro-lectures is not critical, they allow the introduction of topics and ideas that might be useful for the particular dataset being handled in that module. The majority of these micro-lectures are given by the instructor of the course, but each student is required to give one micro-lecture during the semester as well. Requiring students to participate in this way further ties into constructivist theory by encouraging students to create meaning individually and then engage peers in dialogue to increase understanding.

Besides selecting modules and micro-lecture topics, the instructor has additional significant responsibilities throughout the course in their role as the "more skilled member" described by Driver et al. (1994, p. 7). First, the instructor mentors groups as they work on a module. This mentorship can include (but is not limited to) suggesting ideas, pointing out relevant methodology, and facilitating the group work aspect of the course. Second, the instructor works with students to select a topic of their choice to present in their micro-lecture, providing feedback both before and after their presentation. Third, in reviewing the groups' weekly submissions, the instructor is able to provide targeted and timely feedback on the approaches taken that week, giving the groups additional insight into why their chosen methods did (or did not) work.

Course Implementation

The course was taught for the first time at Harvard University in Spring 2014, and again at Simon Fraser University in Spring 2016. The vast majority of the students were seniors concentrating in Statistics, though in both iterations a handful of other majors and early-stage graduate students participated as well. There were several pre-requisites for the course, mostly to encourage sufficient experience with coding and statistical methods. In particular, students were required to have completed a core set of introductory statistics courses as well as two years' worth of courses over all; further, students must have completed at least two courses in applied statistical methods prior to enrolling in this course. With large demand for the course, priority was given to graduating seniors.

The course relied heavily on the data science competition platform Kaggle (www.kaggle.com), which provides educators with an option to run private in-classroom competitions along with all the mechanisms for splitting data, automating student submissions, and displaying and viewing team performance. Effort was made to also have the class participate in public competitions hosted by Kaggle, so students could benchmark against competition outside of the course. Private competitions were used purely due to lack of suitable competitions (for reasons of poor timing, inadequate data, or inappropriate data) on the Kaggle platform that particular semester.

Each meeting of the course started with several micro-lectures (5-10 minute talks on targeted statistical topics), then students broke into their teams to work on their projects. As such, a flexible classroom was necessary, which was achieved by using rolling desks in the first iteration and by seating students at large circular tables in the second iteration.

Method

To study the students' learning as well as their perception of the course, we used a mixed methods approach including a survey for all students and a focus group with a small number of students. The research we conducted was deemed exempt by the research ethics board at Simon Fraser University. While we do not use grades as a measure of overall course success, we do use them to delineate survey results as a function of course performance.

Participants

Our data was collected in the second iteration of the course at Simon Fraser University (Spring 2016). A questionnaire was administered to all twenty-four students in the class, fourteen of whom were female and less than half of whom had English as their first language. Due to its nature as an upper-level undergraduate course, most of the students were in their early twenties, though exact ages are unknown. In addition to the survey, four male students and one female student also participated in a focus group.

Procedure

We mapped the learning outcomes to a survey asking the students to assess how their learning evolved from the beginning to the end of the course. The survey was administered at the courses' completion, and asked students to compare their knowledge pre- and post-course. While the downside of this approach is that it forces students to remember their level at the start of the course, it resolves the issue of students not knowing what they don't know in advance of the course, and also prevents drift in results over the course of the semester. While self-reported measures of efficacy are not a perfect proxy for student learning, they provide insight into the student experience and perceived benefits of participation in the course.

Our data was collected in the second iteration of the course at Simon Fraser University (Spring 2016), using a questionnaire. The questionnaire consisted of eighteen questions asking respondents to read a statement and then rate their agreement with the statement at the beginning and conclusion of the course by selecting one of five options from a balanced scale, ranging from "Strongly Disagree" to "Strongly Agree." To illustrate, the eighteen questions included statements such as "I can explain the benefits of tree-based models relative to linear models," and "I can explain the benefits of version control when writing software." The exact wording for all of the questions can be seen in Figure 2.

To complement the quantitative survey results on student learning, we conducted a student focus group at the end of the course, to collect qualitative data on their learning experience in the class. The focus group was led by one of the authors (Ahrensmeier), who was not an instructor in the course. Audio from the focus group was recorded and later transcribed. The focus group questions addressed the "big picture" of the course. The intention was to take advantage of the participants' experience in "being a student," given that they were in fourth year and about to graduate. These questions aimed at the students' perception of how well the course worked on a day-to-day basis (e.g., potential practical or technical difficulties) and for their education in general.

Data Analysis

We analyzed the survey results in several different ways. First, we compared student responses at the beginning of the class to responses at the end of the class using a Wilcoxon signed rank test (Wilcoxon, 1945). Second, in order to understand how course impacts may have varied for different students, we divided them into two roughly equal-sized groups: those who ended up with an A or A+ grade at the end of the course (which we call the A group), and those who had an A- or lower grade (which we call the non-A group). The A group included 10 people, and the non-A group included 12. Then we determined whether or not question responses differed across these two groups at the beginning and end of the course using a Mann-Whitney test (Mann & Whitney, 1947). Finally, we compared the A and non-A groups by looking at whether or not they showed differences in perceived understanding from the beginning to end of the course using a paired Mann-Whitney test, where a pair consists of a single individual's responses at the beginning and end of the course. The purpose of this comparison is to assess how a students prior knowledge may have impacted their understanding and performance in the course. In order to correct for false-positives due to making multiple comparisons for the 18 survey questions, we utilized a Bonferroni correction to the traditional $\alpha = 0.05$ significance level, which means that for a test result to be considered statistically significant it must be less than $0.05/18 = 0.0028$.

Results

Survey Results

Of the twenty-two students who took the survey, nineteen completed every question in the survey, two did not answer one question in the survey, and one did not answer four of the questions, resulting in exactly one missing data point for six of the questions. With just one exception, every response to the questions asked had students stating they had stayed the same or improved on the given skill. The exception was a student who selected a neutral answer when asked whether or not they knew how to tune a model using cross validation at the beginning of the term and said that they disagreed with this statement at the conclusion of the course. The p-values produced by the Wilcoxon signed rank test comparing student responses for how they rated their skill before and after the course are shown in Figure 1. All of the values in Figure 1 are well below the Bonferroni-corrected $\alpha = 0.05/18 = 0.0028$ significance level, suggesting that for all topics asked about in the questionnaire the class experienced a statistically significant improvement in perceived understanding.

In addition to the p-values showing the differences, we provide histograms showing the distribution of student responses to each of the 18 survey questions in Figure 2. Question 1, which asked the students whether they knew when to apply logistic versus linear regression, exhibited the smallest difference in pre- and post-course rating, but even so, all ten of the students who responded negatively that they knew which method was appropriate prior to taking the course responded with either “Agree” or “Strongly Agree” for the post-course question. Question 6 exhibited the biggest difference in response and of the 15 students who reported that they either disagreed or strongly disagreed that they knew when to apply tree based models prior to the course, ten selected one of the “agree” responses and only one continued to disagree with that statement for the post-course version of the question.

Next we report differences between the A and non-A groups for the pre- and post-course survey questions. Figure 3 shows the p-values from the Mann-Whitney test comparing responses between the two groups for each question, with each arrow indicating how the value changed over the span of the class. In this figure, the value at the tail represents the p-value of the test of differences between how the groups would have answered the question at the beginning of the course and the value at the head of the arrow represents the p-value for how they answered it at the conclusion of the course. With the exception of questions 4 and 7, the p-values increased for all questions, suggesting greater parity between the groups at the conclusion of the class.

Finally, the results of the paired Mann-Whitney test for the A and non-A groups is available in Figure 4. A large p-value in this figure indicates that the students perceived little difference in their understanding of that particular statement from the beginning to the end of the course. For the A group the p-values were largest for questions 1, 5, 14, and 15, which relate to knowing when to apply linear versus logistic regression, level of comfort with applying a linear model to data, working with another statistician to model data, and understanding how best to divide tasks among people with different skill sets, respectively. For the non-A group, the p-values were uniformly small.

Figure 1

P-values for the Wilcoxon Signed-rank Test Comparing Pre- and Post-course Student Skill Ratings in the Post-course Survey

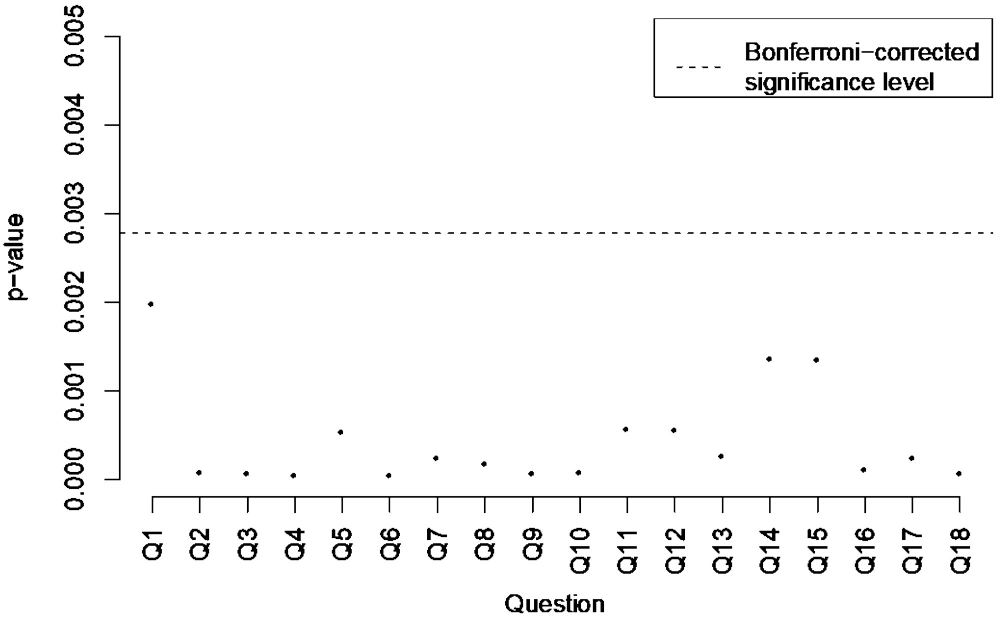
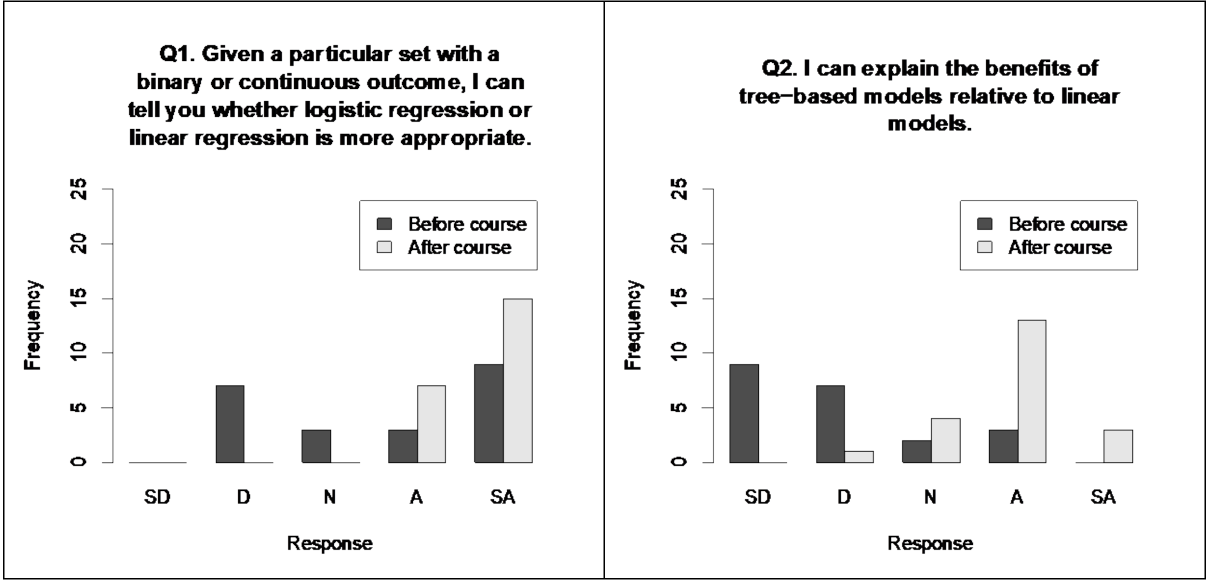
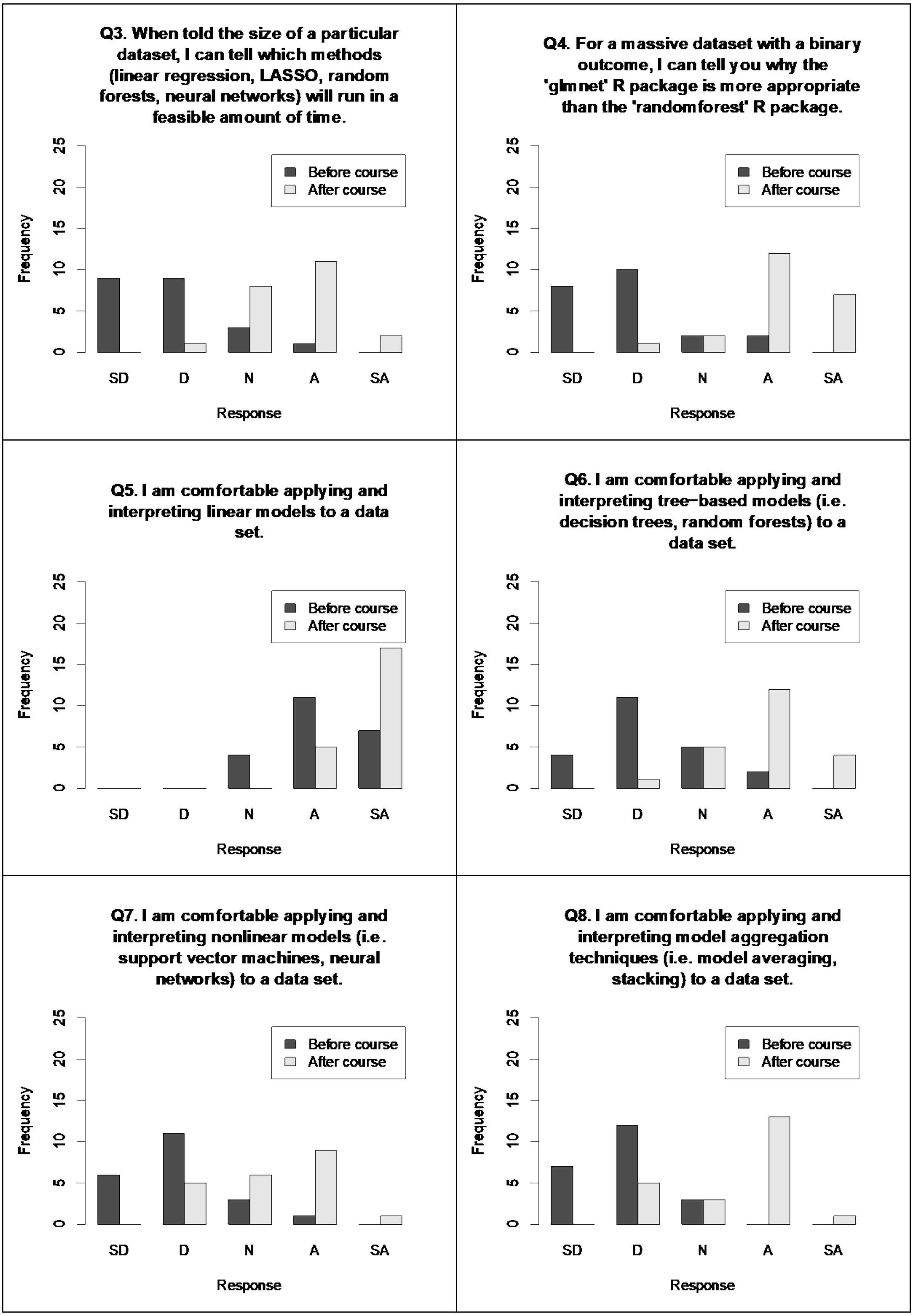
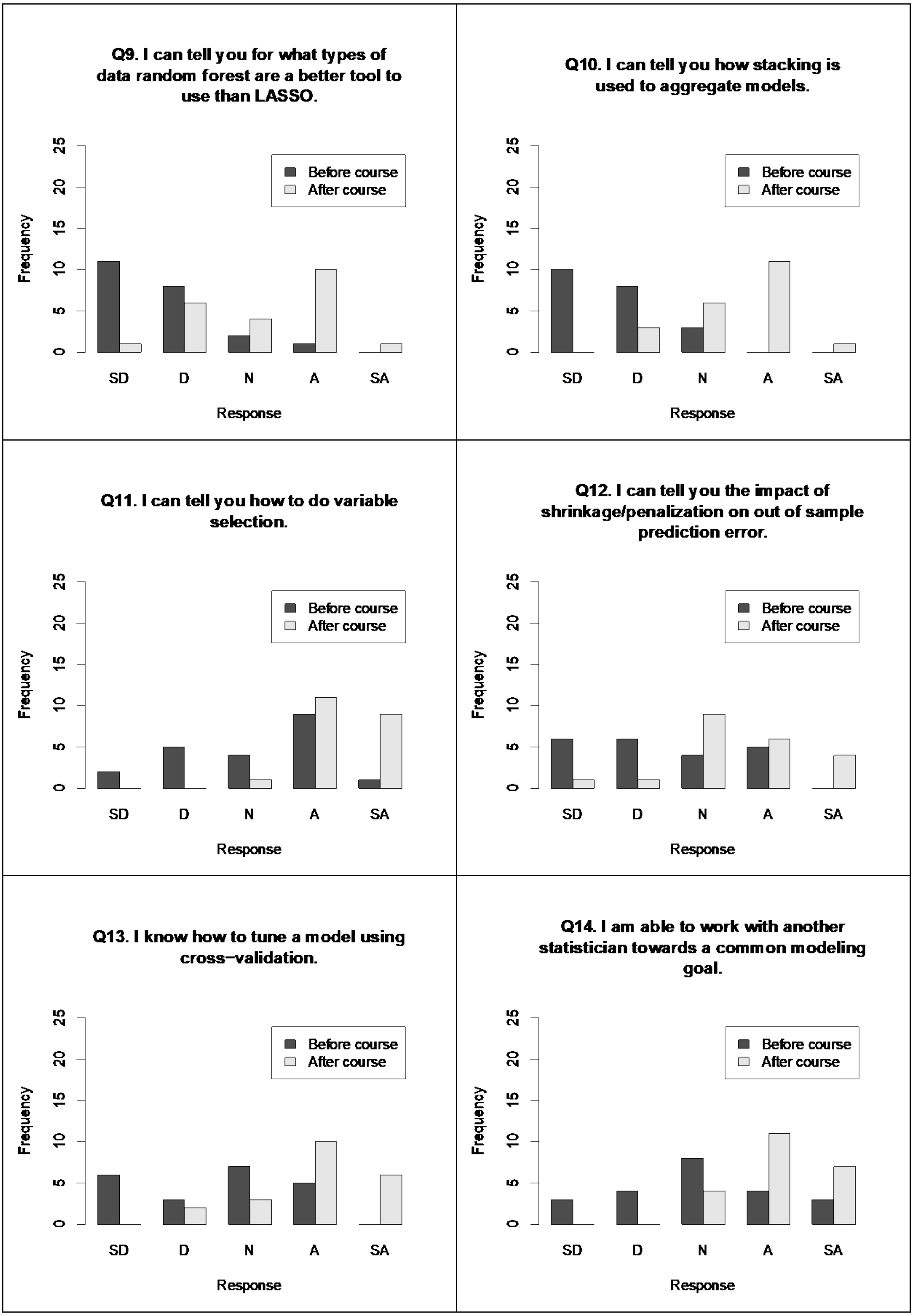


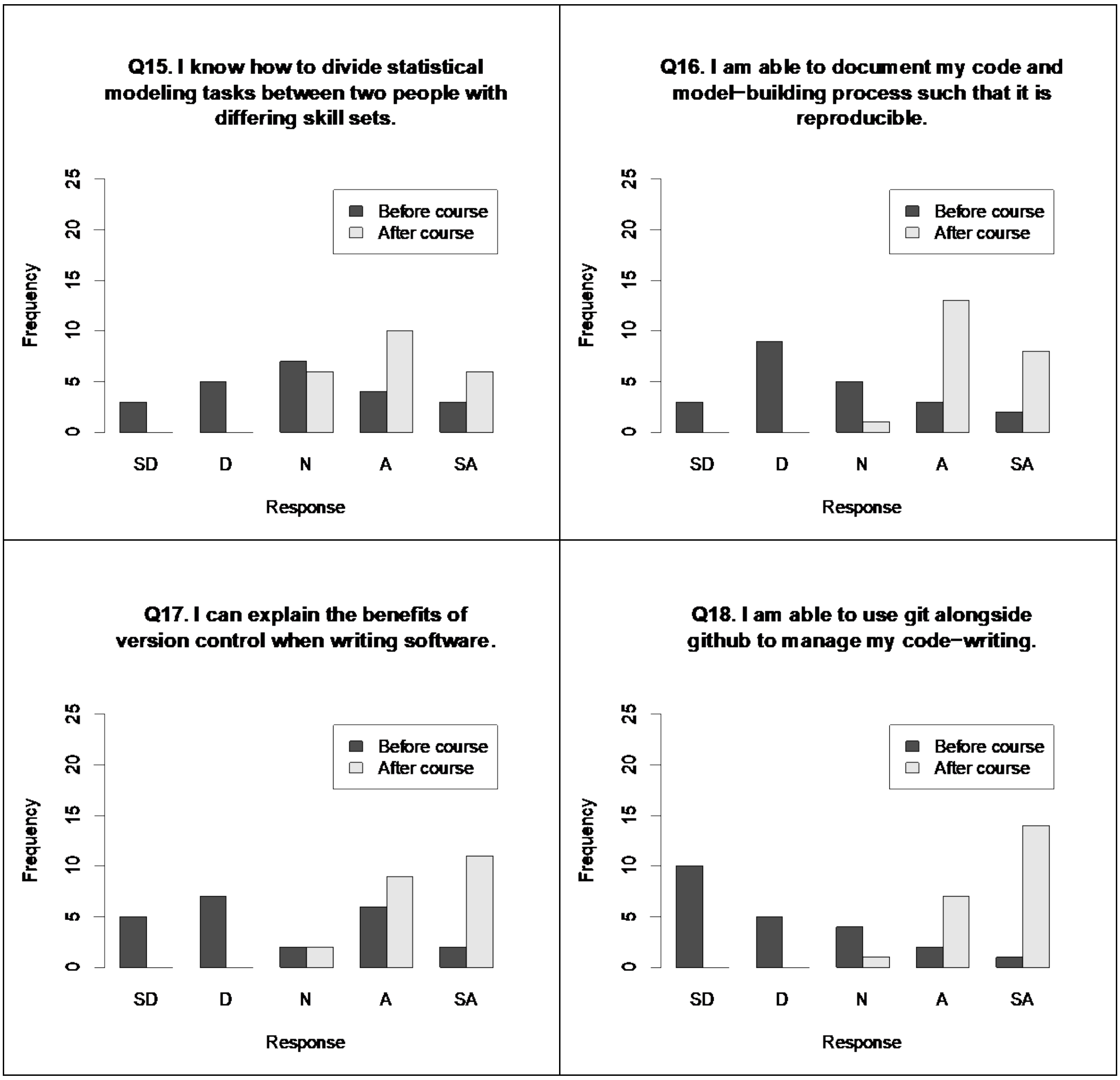
Figure 2

Distribution of Survey Responses for How Students Would Rate their Skill at the Beginning of the Course and How They Would Rate their Skill at the End of the Course




Q3. When told the size of a particular dataset, I can tell which methods (linear regression, LASSO, random forests, neural networks) will run in a feasible amount of time.
Q4. For a massive dataset with a binary outcome, I can tell you why the 'glmnet' R package is more appropriate than the 'randomforest' R package.
Q5. I am comfortable applying and interpreting linear models to a data set.
Q6. I am comfortable applying and interpreting tree–based models (i.e. decision trees, random forests) to a data set.
Q7. I am comfortable applying and interpreting nonlinear models (i.e. support vector machines, neural networks) to a data set.
Q8. I am comfortable applying and interpreting model aggregation techniques (i.e. model averaging, stacking) to a data set.
Before course
After course
Frequency
0
5
10
15
20
25
SD
D
N
A
SA
Response


Q9. I can tell you for what types of data random forest are a better tool to use than LASSO.
Q10. I can tell you how stacking is used to aggregate models.
Q11. I can tell you how to do variable selection.
Q12. I can tell you the impact of shrinkage/penalization on out of sample prediction error.
Q13. I know how to tune a model using cross−validation.
Q14. I am able to work with another statistician towards a common modeling goal.
Frequency
0
5
10
15
20
25
SD
D
N
A
SA
Response
Before course
After course



Q15. I know how to divide statistical modeling tasks between two people with differing skill sets.

Q16. I am able to document my code and model−building process such that it is reproducible.

Q17. I can explain the benefits of version control when writing software.

Q18. I am able to use git alongside github to manage my code−writing.

Figure 3

P-values from a Mann-Whitney Test Comparing Question Response Differences Between the A and Non-A Groups

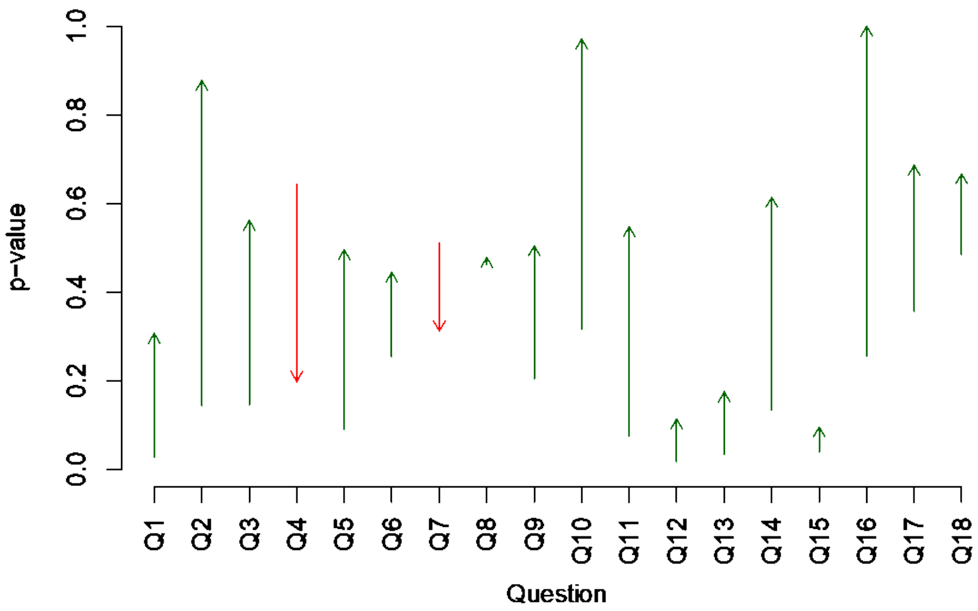
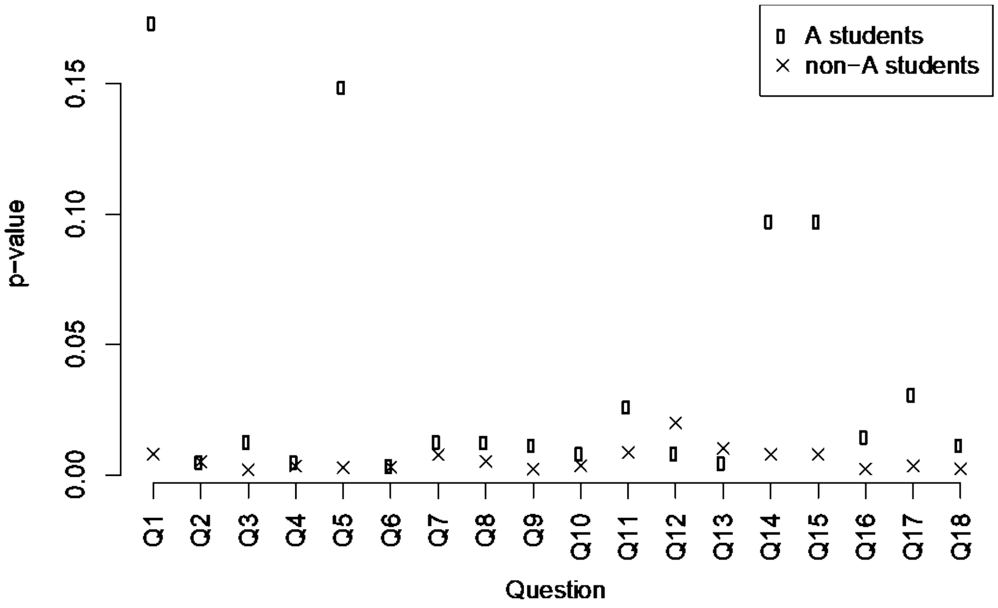


Figure 4

P-values for Mann-Whitney Test Comparing Student Ratings of their Skill Prior to and After the Course in the Post-course Survey, Divided by A and Non-A Groups



Focus Group

A summary of the questions asked in the focus group and the student responses is provided in the following paragraphs.

1. *What is the first thing you would want people to know about this course?*

"It was exciting!" was the first answer. The student explained that what made the course exciting was the use of "interesting data, really really good data," which motivated them to put in more work than for a normal course. In the words of other students, "it makes you want to learn more and more," as opposed to other courses in which the instructor asks them to memorize knowledge, and "the only limit in this course is your imagination, you can do all the things you want," very fancy things or just basic algorithms. While all students agreed that the course required a lot of work, it was worth it for most of them, because "I feel way more confident" [in analyzing data], as one of them put it (reflecting learning goal 2, ability to apply a broad suite of tools to data science problems).

Aside from being motivated by the content, the students pointed out that they were driven by "competitiveness." One student repeated, half jokingly, that he always wanted to beat L. (another student), and they agreed that they all "just wanted to impress the instructor," the professor. At the same time, they pointed out the "welcoming environment" in the course and the impression that "both instructors [professor and TA] care for you to do well."

2. *What worked (particularly) well for you? Why?*

"The sharing of code worked very well for me" was one response, but the student also mentioned that "I have a strong comp sci background" (see the next question for a different view). Other students mentioned "you look at other people's code and get ideas", or learning about "all these different things like packages I hadn't seen before" and "trying them out". Another student specifically pointed out the opportunity to "work in class", a dedicated time to work on their project in an environment similar to a lab, where they can ask the instructors for help. Aside from the emphasis on practice, these responses reflect on learning goal 3, collaboration with other statisticians. However, it should be noted that several students mentioned that they preferred to work on their own when possible.

3. *What did not work so well? Any suggestions for improvement?*

The students identified three areas for improvement. One of them is the wide variety in students' coding background. They emphasized that experience with applying R is not sufficient, because even with that background they were "playing catch up a lot." Students with little coding experience also did not benefit as much from other students' code they were supposed to improve, because they did not understand well enough how it worked. For this specific issue, they suggested to provide more explanations with the shared code, which may further address learning goal 4. In general, they suggested to include a stronger coding requirement in the prerequisites for the course, possibly using python.

The second area for improvement they identified was the grading scheme. As one student thoughtfully expressed it, they identified the tension between "what works for real life" vs "what works for learning": After the first two weeks of a module, most students would focus on the one best performing model trying to improve it further, which—they think—restricted their options for learning: The "code is shared for people who get stuck, but if someone else has already optimized it, you don't learn from it, you learn more if you fiddle with your own code because there is still room to improve." One student mentioned that although creativity was encouraged in the course, they felt it was not necessarily rewarded by the grading scheme. Another one described a situation in which they spent many hours trying to improve one model that did not perform that well, and in the end received a lower mark than a student who copied and submitted existing code (without improving it) that already performed better. To address this imbalance, the students suggested to use more components for grading than only the ranking. This part of the student feedback is related to learning goal 1, ability to identify a set of suitable (and rule out unsuitable) methods for the task at hand. Does the students' desire to improve their technical skills conflict with the goal to focus on finding the best method for solving a problem? On the other hand, the responses illustrate that students are comfortable experimenting with various statistical methods (learning goal 5).

The third area for which the students mentioned some issues was the group work. They appreciated that the instructor made suggestions on how to split the work, but that in the end, it "completely depended on who you teamed up with." Given that this problem is by no means specific to this course, it is not surprising that the students were unable to suggest a solution. They did point out that "a lot of people were willing to do the work," as should be expected in a fourth year class, but still "some people coast."

4. *Would you like to see more courses taught this way? Or would you like to see specific aspects of this course in other courses?*

One student suggested that it would be nice to have a course like this at each level. The other students specifically pointed out the "data-driven approach" with "non-perfect data" that they would like to see in other courses, and the opportunity to "experiment" in class.

Discussion

As shown in Figure 1, all of the p-values are statistically significant, suggesting that students felt the class increased their understanding. The most significant improvement was in regards to how to use and interpret tree-based models, followed closely by the related topic of when it is more appropriate to use the glmnet R package or the randomforest package. The question that showed the least significant difference in response values asked about whether or not the respondent understood when logistic regression is a better modeling choice than linear regression. This is unsurprising, as this is a topic that most, if not all, of the students should have encountered in their prerequisite courses. It is worth noting, however, that the p-value associated with this question is still statistically significant, indicating that students experienced increased confidence with even familiar statistical concepts while enrolled in this course.

In addition to the statistical significance represented by the p-values, we can gain a sense of practical significance by considering how many students rated their skills differently prior to and after the course and in what way. Figure 2 shows how the distribution of responses varies for the pre- and post-course questions. Regardless of whether the question was one that students felt relatively confident about at the outset of the course (such as questions 1 or 14) or one with which they were unfamiliar (such as question 6), responses to the questions show marked improvement across the board when comparing answers for the pre- and post-course questions in the survey.

Next, we turn to comparison of the A and non-A groups. In the pre-course questions, differences between the two groups resulted in the lowest p-values for questions 1, 12, 13, and 15. Although these values are larger than the Bonferroni corrected α, they are all less than 0.05 and taken in concert still provide some evidence of group differences at the start of the course. In the responses to the post-course questions there were no significant differences between the A group and non-A group for any of the questions and the p-values for nearly all questions increased in size, implying that the question responses between A and non-A groups were even more indistinguishable after completing the curriculum.

The two questions that experienced a decrease in p-value were questions 4 and 7, dealing with the glmnet and randomforest R packages, and the application and interpretation of nonlinear models, respectively. In both cases, responses to the pre-course questions ranged between "Strongly Disagree" and "Neither Agree nor Disagree," and the distribution of responses was very similar between the A and non-A groups. When compared to the post-course versions of these questions, the A group responses fell primarily into the "Agree" and "Strongly Agree" categories, while the non-A group responses included some "Disagree" and "Neither Agree nor Disagree" answers. Thus, these questions appear to show that both groups had little prior knowledge about these topics, but that at the conclusion of the course the A group felt more confidence in their mastery of these particular topics.

The results of the paired Mann-Whitney test in Figure 4 provides more insight into differences in perceived understanding at the beginning of the course between the A and non-A group. The A group produces p-values less than α for none of the questions while the non-A group produces p-values less than α for five of the eighteen. The most obvious differences between the two groups are with question 1, about when to use logistic versus linear regression, and question 5, dealing with the application and interpretation of linear models. This is primarily due to the fact that the majority of the students in the A-group were already familiar with the concepts prior to the start of the course, with all but two of the students indicating that they either agreed or strongly agreed with the statement in question 5 and all but 3 stating that they agreed or strongly agreed with the statement in the first survey question. Questions 14 and 15, both of which concern working with other people, exhibit the next largest differences between the A and non-A groups. Though the number of "Agree" and "Strongly Agree" responses are not as overwhelming as in questions 1 and 5, the larger p-values for the A group can be explained by the same logic. In other words, some of the lack of improvement in confidence demonstrated by the A-group is due to a meaningful amount of prior learning.

Overall, the qualitative data from the focus group complement the quantitative survey data in illustrating how well the learning outcomes were achieved. Beyond this triangulation, the student responses also serve to tell an interesting story about their transformative and rewarding experience in this course, and their thoughtful reflections on their own learning. Generally, their comments suggest that the course provided a constructive learning environment, with particular emphasis on the applied nature, instructor availability, and the peer learning aspect. They also provided several valuable suggestions for improving the course, particularly, rewarding creativity through the grading scheme in order to encourage more diverse solutions within each module and requiring more coding experience prior to enrolling in this course.

The student feedback obtained in the focus group and from the survey are consistent with the findings of other studies assessing the efficacy of a data-first approach to learning. For example, Fawcett (2017) observed improved student performance among students who used real-life data sets and Lazar et al. (2011) found that students expressed that their capstone course "prepare[d] them for the real world and [gave] them confidence that they will be able to handle data analysis in a more unstructured environment" (p. 189). These sentiments are echoed in the feedback from the focus group and in the survey responses here. In addition to the data-first benefits identified by Lazar, Fawcett, and others, we found that the novel aspects of this course had a positive effect. Namely, students were motivated by the competition aspect of the course and sharing of analyses allowed students to learn from and construct knowledge along with their peers.

Conclusion

In this paper we presented a new course design implemented over two iterations at two different universities. The course changes the normal approach of incorporating data into statistical education, structuring around curated datasets (with associated problems and objective functions) and allowing students the flexibility to choose, apply, and compare different methods. By having students submit their work for evaluation every week and sharing their write-up and code with each other, the course operates much like a genetic algorithm whereby each week low performing students leave behind their unsuccessful methods to adopt the methodology of the high performing groups, and subsequently build and expand the methods from there.

In the second delivery of the course, a post-course survey was conducted, which demonstrates that students felt that learning outcomes were achieved and that the course narrowed the perceived gap between low-performing and high-performing students over its duration. Specifically, the high-performing students (as measured by their final cumulative grade) self-rated their skills at the beginning of the course more highly than their lower-performing peers, with this difference disappearing for questions asking them to rate their skill at the conclusion of the course, indicating that the lower-performing students tended to close the gap on their peers over the length of the course.

Because the data in this paper was collected on a specialized group of students, we emphasize that more work needs to be done to determine whether or not this type of approach is globally effective across disciplines or for learning other types of statistical material. That being said, we believe that elements of the structure presented here are worth exploring for any course with an applied analysis component. Given that the demand for the course has so far exceeded the available student space, with a lottery determining enrollment, we hope in the future to further study the outcomes of those who were randomly selected for the course vs. those who were not.

As a course which requires students to select and apply methods of their choice to large, complex datasets, this course is intended as a senior undergraduate or graduate course. However, akin to how many machine learning courses might end with a Kaggle competition amongst students, elements of the course design could easily be borrowed as an element of another course. Courses on statistical communication, in particular, could straightforwardly be merged to teach both problem-solving and the communication thereof in the same structure. As a final comment, those attempting to apply these ideas should be aware of the challenges the course creates for instructors. In particular, given the unique structure of the course, with new methods being applied daily, the course requires an instructor with a broad base of knowledge in statistical models, as well as experience applying these methods to real data.

References

Ashford-Rowe, K., Herrington, J., & Brown, C. (2014). Establishing the critical elements that determine authentic assessment. *Assessment & Evaluation in Higher Education*, *39*(2), 205-222. https://doi.org/10.1080/02602938.2013.819566

Bond, M. E., Perkins, S. N., & Ramirez, C. (2012). Students’ perceptions of statistics: An exploration of attitudes, conceptualizations, and content knowledge of statistics. *Statistics Education Research Journal*, *11*(2), 6-25. https://iase-web.org/documents/SERJ/SERJ11(2)_Bond.pdf?1402525003

Cobb, P., & McClain, K. (2004). Principles of instructional design for supporting the development of students' statistical reasoning. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 375-395). Springer. https://doi.org/10.1007/1-4020-2278-6

Dierdorp, A., Bakker, A., Eijkelhof, H., & van Maanen, J. (2011). Authentic practices as contexts for learning to draw inferences beyond correlated data. *Mathematical Thinking and Learning, 13*(1-2), 132-151. https://doi.org/10.1080/10986065.2011.538294

Driver, R., Asoko, H., Leach, J., Mortimer, E., & Scott, P. (1994). Constructing scientific knowledge in the classroom. *Educational Researcher, 23*(7), 5-12. https://doi.org/10.3102/0013189X023007005

Duit, R. (1996). The constructivist view in science education – what it has to offer and what should not be expected from it. *Investigações em Ensino de Ciências, 1*(1), 40-75. https://doaj.org/article/364ca28aa72940c3b0689d0bf435401a

Dunn, D. S., Smith, R. A., & Beins, B. C. (Eds.). (2012). *Best practices in teaching statistics and research methods in the behavioral sciences*. Routledge.

Fawcett, L. (2017). The case project: Evaluation of case-based approaches to learning and teaching in statistics service courses. *Journal of Statistics Education, 25*(2), 79-89. https://doi.org/10.1080/10691898.2017.1341286

Garfield, J., & Ben-Zvi, D. (2009). Helping students develop statistical reasoning: Implementing a statistical reasoning learning environment. *Teaching Statistics, 31*(3), 72-77. https://doi.org/10.1111/j.1467-9639.2009.00363.x

Holland, J. H. (1984). Genetic algorithms and adaptation. In O. G. Selfridge, E. L. Rissland, & M. A. Arbib (Eds.), *Adaptive control of ill-defined systems. NATO conference series (II systems science), vol 16.* Springer US. https://doi.org/10.1007/978-1-4684-8941-5_21

Lazar, N. A., Reeves, J., & Franklin, C. (2011). A capstone course for undergraduate statistics majors. *American Statistician, 65*(3), 183-189. https://www.jstor.org/stable/24591414

Makar, K., & Ben-Zvi, D. (2011). The role of context in developing reasoning about informal statistical inference. *Mathematical Thinking and Learning, 13*(1-2), 1-4. https://doi.org/10.1080/10986065.2011.538291

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 50-60.

Neumann, D. L., Hood, M., & Neumann, M. (2013). Using real-life data when teaching statistics: Student perceptions of this strategy in an introductory statistics course. *Statistics Education Research Journal*, 12(2), 59-70. http://iase-web.org/documents/SERJ/SERJ12(2)_Neumann.pdf

Pfannkuch, M. (2011). The role of context in developing informal statistical inferential reasoning: A classroom study. *Mathematical Thinking and Learning* 13(1-2), 27-46. https://doi.org/10.1080/10986065.2011.538302

Scheaffer, R. L. (2001). Statistics education: Perusing the past, embracing the present, and charting the future. *Newsletter for the Section on Statistical Education of the American Statistical Association, 7*(1), 2-9.

Singer, J. D., & Willett, J. B. (1990). Improving the teaching of applied statistics: Putting the data back into data analysis. *The American Statistician, 44*(3), 223–230. https://doi.org/10.1080/00031305.1990.10475726

Stander, J., & Dalla Valle, L. (2017). On enthusing students about big data and social media visualization and analysis using R, RStudio, and RMarkdown. *Journal of Statistics Education, 25*(2), 60-67. https://doi.org/10.1080/10691898.2017.1322474

Tobin, K. G. (Ed.). (1995). *The practice of constructivism in science education*. Routledge. https://doi.org/10.4324/9780203053409

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin, 1*(6), 80-83. https://www.jstor.org/stable/3001968